Queries and tags in affect-based multimedia retrieval

Kierkels, Joep Johannes Maria; Soleymani, Mohammad; Pun, Thierry

QUERIES AND TAGS IN AFFECT-BASED MULTIMEDIA RETRIEVAL

Joep J.M. Kierkels, Mohammad Soleymani, Thierry Pun

University of Geneva, Computer Science Department
Battelle Building A, 7 Route de Drize
CH - 1227 Carouge, Geneva, Switzerland

ABSTRACT

An approach for implementing affective information as tags for multimedia content indexing and retrieval is presented. The approach can be used for implicit as well as explicit tags and is presented here using data recorded during the viewing of movie fragments containing annotations and physiological signal recordings. For retrieval based on affective queries, a representation of the query-words is defined in the arousal-valence space in the form of a Gaussian probability distribution and a retrieval method based on this representation is presented. Validation of retrieval accuracy is performed using Precision and Recall parameters. Results show that the use of arousal and valence as affective tags can improve retrieval results.

Index Terms— Tagging, Indexing and Retrieval, Emotions, Arousal, Valence

1. INTRODUCTION

Tags attached to multimedia items such as music, video and audio clips greatly facilitate the speed and accuracy at which a search in a large multimedia database can be performed. In order to have such tags available, there is a need for methods that attach tags to the multimedia items [5;9].Implicit tagging of items is the process of creating tags in which, as opposed to explicit tagging, a tagger does not invest effort into the creation of tags but rather a tag is generated automatically based on automated observations and analysis of the multimedia item and possibly of the tagger. Implicit tagging has been receiving a growing amount of interest over the past years e.g., [7], especially since the amount of multimedia content is growing ever larger and one should therefore minimize the effort needed to tag and categorize content. Among the possible tags that can be attached to a multimedia item, some tags refer to more subjective and emotional properties of an item; these are the affective tags [6]. Implicit affective tagging thus refers to the effortless generation of subjective and/or emotional tags.

While watching pictures or video clips or listening to music, a user may experience certain feelings and emotions [3;4;11] which are reflected in changes in 1) physiological signals, e.g., sweating, 2) changes in facial features, e.g., frowning, and 3) changes in vocal features, e.g., laughing. If an implicit affective tagging algorithm thus monitors the reactions of a person while he or she watches a specific multimedia item, this information can be used to continuously and effortlessly generate affective tags.

When using implicit affective tags, there often is a discrepancy between how content is commonly searched for, e.g., using queries as “scary” or “romantic”, and how affect is being detected, e.g., using physiological signal analysis.

In order to create a meaningful linkage between this search vocabulary of query words and the affect monitoring it is worth considering three possible options:

1. The signal(s) which is recorded as part of the affect monitoring is transformed to a specific tag that corresponds exactly to the search vocabulary, e.g., “romantic”. The key advantage of this option is that search query retrieval is very fast because the tags are exact matches to the queries. The main disadvantage of this approach is the limitation that is put on retrieval. Only certain specific searches (corresponding to the labels) can be performed and there will be no gradation in data which have been assigned with the same tag.
2. The signal(s) which is recorded as part of the affect monitoring as a whole is stored as the tag. The sole advantage of this option is the fact that no (possibly) relevant information is lost. The big disadvantages clearly are the size of the tags and the fact the every query would first need to process the stored tags before returning the results on a query.
3. The signal(s) which is recorded as part of the affect monitoring is transformed to one or more continuous tagging quantities which span a continuous space. A search query initializes a transformation of the query word to this space and searches in a region of the space corresponding to the given query word. This results in a fast retrieval method with the possibility to distinguish between closely related data items. Moreover, the search vocabulary can be extended without the need to redo tagging. The concept of having quantities spanning a continuous space related to emotional feelings is well known from emotion literature. Examples of such are arousal, valence, control and unpredictability [2].

In this work we will show how this third option, with no real disadvantages compared to the other two options and illustrated in Figure 1, can be implemented in such a way that personalised retrieval and multi-term queries are feasible. The arousal- and valence- (AV) space [8] will be used to represent affect because many modalities for affect detection have been shown to project onto the AV space (e.g., [3;10]) and the use of this space thus allows for cross-modality tagging. Although results will be shown for both explicit affective tags and implicit affective tags, the emphasis in the current work is thus on tag-based retrieval rather than tag generation.

For retrieval, it will first be shown how specific query words project to an area inside the AV space and how multi-term queries can be implemented. Next, it will be shown how items in a database can be sorted according to their AV values based on such an area inside the AV. Finally, the accuracy of the sorted list will be validated using precision and recall measures. Sorting accuracy (using either explicit or implicit tags) will be compared against both a classical sorting method and a randomized sorting.

Examples will be worked out using data from a study in which affective tags for short video clips were generated both explicitly and implicitly.

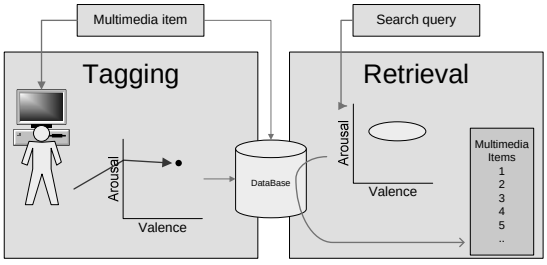

Figure 1: Illustration of the steps involved in tagging of multimedia items and retrieval based on a query.

When tagging content, only the AV values are stored as affective tags. To achieve personalized retrieval, characteristics of users and taggers can be exploited. If available, tagger characteristics such as age, gender and nationality, can be added as context tags. When retrieving content, only affective tags of taggers with similar characteristics are used.

2. METHODS

Seven adults participated in the experiments and watched 64 short video clips taken from 8 different movies. Prior to watching these clips, the concept of AV was explained to the participants. Clips were watched consecutively in a randomized sequence, this was done in 2 sessions of 32 clips each in order to reduce the uninterrupted on-task time of the participants. Each clip lasted approximately two minutes and following each clip, the participants were asked to assess their AV-values related to this clip and to write down one or more keywords that describe their feeling towards it (*explicit tagging*). Following this self-assessment, a neutral clip was shown for 30 s, followed by the next short video clip. Experiments are described in details in [10]. As nationality differed between participants, tags were given in different languages. All tags were converted afterwards to English. While participants watched the clips, we recorded their electrocardiogram (ECG), galvanic skin response (GSR), respiration, and skin temperature using the BioSemi active 2 system and standard BioSemi peripheral sensors (*http://www.biosemi.com/*).

2.1. Implicit tagging

Estimation of AV values based on recordings of physiological data is well documented in literature and the proposed retrieval-method merely requires that these values are stored properly.

As explained in [10], several features are extracted from the recorded physiological signals and AV values are estimated based on a linear combination of these features (*implicit tagging*). For this estimating procedure it is required to include the self assessed AV values.

2.2. Retrieval

2.2.1. Converting query words to probability distributions

As illustrated in Figure 1, retrieval starts by defining how a query-word can be converted into something that has a meaning in the AV space. The keywords given by all participants are assumed to be representatives of the vocabulary of search queries.

For a given keyword, the corresponding AV values usually roughly cluster in a certain area of the AV space, as is illustrated in Figure 2. Figure 2B is included to illustrate personalized retrieval. The mean arousal is different (one-way ANOVA ($F=11.1$, $p<0.005$)) for the characteristic of French vs. non-French speaking and will lead to retrieval of different video clips.

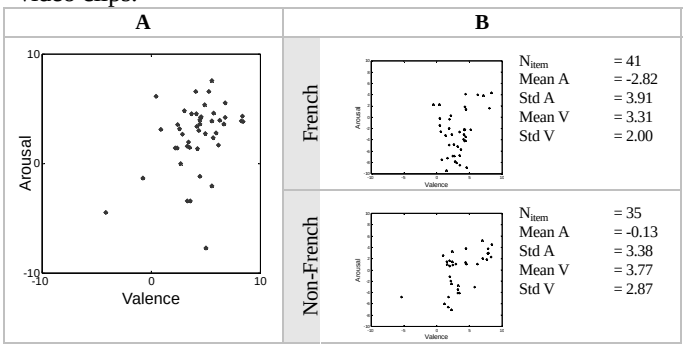

Figure 2: A: Self-assessed AV values for clips tagged 'joyful'. B: Self-assessed AV values for 'amused' clips by either French or non-French speaking participants.

The morphology of such clusters has not (yet) received much attention in research and is often represented as an ellipsoid, e.g., [8;12]. Because it is not the intention here to study this morphology and because apparently the use of ellipsoids is intuitively correct to viewers, it will be assumed that these clusters originate from a 2D Gaussian probability distribution,

the contour-lines of which resemble ellipsoids. The probability distributions for a keyword can be derived as follows:

- All AV values belonging to assessments with a specific keyword, *k*, (e.g., for 'joyful' all points shown in Figure 2A) are grouped in data matrix $\underline{\underline{X}}$ (N_kx2), with N_k the number of times that keyword *k* occurs.
- The mean AV values are subtracted from $\underline{\underline{X}}$ leading to zero-centred matrix $\underline{\underline{X}}_0$.
- Using principal components analysis, two perpendicular axes of maximal variance are determined. For each *k*, a matrix $\underline{\underline{W}}$ (2x2) defines the projection of $\underline{\underline{X}}_0$ in the AV space to a cluster $\underline{\underline{Y}}$ in the new PCA space as

$$\underline{\underline{Y}}^k = \underline{\underline{X}}_0^k \cdot \underline{\underline{W}}^k. \quad (1)$$

- Two variances are determined from

$$[v_{k1}, v_{k2}] = diag\left(\left(\underline{\underline{Y}}^k\right)' \cdot \left(\underline{\underline{Y}}^k\right)\right). \quad (2)$$

- Using v_{k1} and v_{k2}, and assuming a Gaussian distribution, the probability that a point $\underline{x}$ (1x2) in the AV space has the annotation *k* can be derived from

$$p\left(k \mid \underline{x}\right) = \frac{1}{2\pi\sqrt{(v_{k1} \cdot v_{k2})}} \cdot e^{-\left(\frac{y(1)^2}{2 \cdot v_{k1}} + \frac{y(2)^2}{2 \cdot v_{k2}}\right)}. \quad (3)$$

In Eq. (3), y(1) and y(2) are elements of $\underline{y}$ (1x2) which is derived from $\underline{x}$ in a way similar to Eq. (1). To exclude statistically insignificant keywords, only keywords with $N_k > 15$ are included in this study.

2.2.2. Ranking a database according to a query

In the database of 64 clips, each clip has seven AV tags attached to it, one from each participant. For each of these AV tags, the probability $p\left(k \mid \underline{x}\right)$ is computed and the averaged probability *P*(*k*) that this clip would receive the annotation *k* is determined from

$$P(k) = \frac{1}{7}\sum_{i=1}^{7} p(k \mid \underline{x}). \quad (4)$$

Sorting clips according to their *P*(*k*) will sort the database according to the relevance of the query. Using multi-term queries merely requires that averaged probabilities for separate terms are multiplied, $P(k_1,k_2) = P(k_1)\,P(k_2)$.

2.3. Validating sorted databases

Once all clips in the database are sorted according to Eq. (4) for a specific query, sorting accuracy can be determined. Because the self-assessed keywords were not used in the sorting process (Section 2.2.2), these keywords can be used as indicators of the accuracy of the sorting process. However, self-assessed keywords were used in the process of converting query words to probability distributions (Section 2.2.1). To avoid any bias in results because of this, the data of seven participants are divided into two sets. A set of training data for query conversion, which contains annotations of four participants only, and a set of test data for the ranking of clips containing annotations of the remaining three participants. In this way the sorting of the clips is independent of the keywords given for the test data.

To quantify the accuracy of the sorting of clips, recall and precision [1] are computed as

$$\begin{aligned} precision &= \left|\{relevant\} \cap \{retrieved\}\right| / \left|\{retrieved\}\right| \\ recall &= \left|\{relevant\} \cap \{retrieved\}\right| / \left|\{relevant\}\right| \end{aligned}. \quad (5)$$

The set of *relevant* clips represents those clips that should be retrieved because at least one of the participants in the test group labeled the clip with the query word. The set of *retrieved* clips is filled iteratively. Initially, only the clip with the highest probability *P*(*k*) is in this set and precision and recall are computed. Subsequently the clip with the second highest *P*(*k*) is added to the set and precision and recall are computed again, followed by the third highest etc. If 64 clips are added to the retrieved list, recall will be one.

Precision and recall of the sorted clips will be compared to precision and recall computed over a randomly sorted list and over a list sorted according to a classical sorting method. The classical sorting method sorts the clips based on the number of times a clip received the specific keyword from the participants in the training set.

3. RESULTS & DISCUSSION

Figure 3 shows the precision and recall for the sorted lists. The shown graphs are averaged over all keywords and over all possible (=35) combinations of dividing the seven participants over the test and training sets.

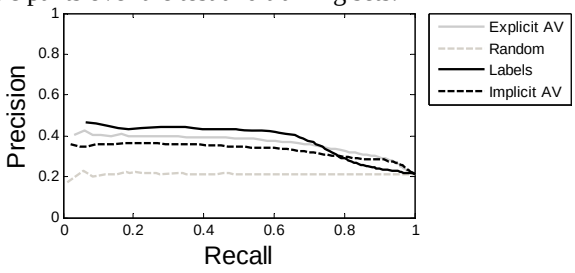

Figure 3: Precision-Recall curves illustrating the accuracy of retrieval for the four different sorting methods.

The solid gray line is based on the self-assessed (explicit) AV values; the dashed black line is derived from physiological signals (implicit). The desired result would have high precision for all possible recall values, indicating that the chance of retrieving a relevant clip in the next iteration (described in Section 2.3) is always high. This figure shows that

1. The three sorting methods have better results than random sorting.
2. The random sorting has a constant precision level, indicating that the chance that the next item on the list

will be relevant is identical for the beginning and the end of the list.

3. The three other methods have decreasing precision when recall increases. This is to be expected because high recall implies that many relevant clips have already been retrieved and the ones which were not yet retrieved are those which are more difficult to retrieve, e.g., because the AV for this clip's keyword is not as strong (or stronger) than average.
4. The results based on explicit AV values are continuously better than the results based on implicit AV values, which could reflect the fact that implicit tagging requires an extra step in which the AV values are derived from physiological features. Because of this step, implicit AV values may be less accurate. However, it should be remembered that the implicit AV values can be determined without asking for a participant's feedback which greatly increases it's usability in practical setups.
5. The classical sorting method performs best for low recall values. Thus, if all clips are appropriately labeled with keywords, retrieval based on these keywords will have the highest probability of retrieving relevant clips among the firstly retrieved clip. A non-exhaustive search in which only a limited number of relevant items needs to be retrieved is thus best performed using classical sorting.
6. The classical sorting method performs less for higher recall values. Thus, it is less successful in performing an exhaustive search and retrieving all relevant clips.

It should be noted that all self-assessments are considered to be accurate and all reported values and keywords are used throughout the study. This relates to our believe that the best judge of someone's emotion is that person himself. Nevertheless, once more data are available, outlier-detection could be implemented.

4. CONCLUSIONS

In this paper, an approach for the retrieval of multimedia items is proposed that is based on affective labeling using the arousal-valence space. It can be implemented using either explicitly or implicitly obtained labels. From the results, it can be derived that performance of the proposed approach is similar to classical retrieval based on keywords. Accuracy is slightly lower in a non-exhaustive search, but slightly higher in an exhaustive search. Results further show that the retrieval based on implicitly obtained arousal and valence (from physiological signals), is slightly less accurate than based on explicit tags. It was also demonstrated how the approach can be used for multi-term queries and for personalized retrieval.

5. ACKNOWLEDGEMENT

This work was supported by the EU Network of Excellence Petamedia.

6. REFERENCES

[1] M. Buckland and F. Gey, "The Relationship Between Recall and Precision," *Journal of the American Society for Information Science*, vol. 45, no. 1, pp. 12-19, 1994.

[2] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050-1057, 2007.

[3] P. Gomez and B. Danuser, "Relationships between musical structure and psychophysiological measures of emotion," *Emotion*, vol. 7, no. 2, pp. 377-387, 2007.

[4] P. Gomez, W. A. Stahel, and B. Danuser, "Respiratory responses during affective picture viewing," *Biological Psychology*, vol. 67, no. 3, pp. 359-373, 2004.

[5] A. Hanjalic, R. Lienhart, W. Y. Ma, and J. R. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?," *Proceedings of the Ieee*, vol. 96, no. 4, pp. 541-547, 2008.

[6] M. E. I. Kipp, "@toread and Cool: Subjective, Affective and Associative Factors in Tagging," *Proceedings of the 36th annual conference of the CAIS*, Vancouver, June 5-7, 2008.

[7] A. Nijholt, "Girlfriends and Strawberry Jam: Tagging Memories, Experiences, and Events," *PETAMEDIA 'Implicit Tagging' Workshop*, London, September 5, 2008.

[8] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, pp. 715-734, 2005.

[9] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks, "Multimedia indexing through multi-source and multi-language information extraction: the MUMIS project," *Data & Knowledge Engineering*, vol. 48, no. 2, pp. 247-264, 2004.

[10] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses," *IEEE International Symposium on Multimedia*, Berkeley, December 15-17, 2008.

[11] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "International Conference on Multimodal Interfaces," *Proceedings of the 10th international conference on Multimodal* interfaces, Chania, October 20-22, 2008.

[12] C. M. Whissell, "The dictionary of affect in language," in *EMOTION Theory, Research and Experience, Vol. 4, The Measurement of Emotions*. R. Plutchik and H. Kellerman, Eds. Academic press, 1989, pp. 113-131.